
High Throughput Sequencing
Algorithms and Applications

*A special track of the ISMB 2020 meeting*
Montreal, Canada, July 15-16, 2020

# ISMB 2020 HiTSeq Track Proceedings

Montreal, Canada (virtual)
July 15-16, 2020
http://www.hitseq.org

**Organizers:**

Can Alkan
Bilkent University, Bilkent, Ankara, Turkey
E-mail: calkan@gmail.com

Ana Conesa
University of Florida, Gainesville, Florida, USA
E-mail: vickycoce@gmail.com

Francisco M. De La Vega, D.Sc.
Stanford University, and TOMA Biosciences, USA.
E-mail: Francisco.DeLaVega@stanford.edu

Dirk Evers
Dr. Dirk Evers Consulting, Heidelberg, Germany
E-mail: dirk.evers@gmail.com

Kjong Lehmann
ETH-Zürich, Zürich, Switzerland
E-mail: kjong.lehmann@inf.ethz.ch

Layla Oesper
Carleton College, Northfield, MN, United States
E-mail: loesper@carleton.edu

Gunnar Rätsch
ETH-Zürich, Zürich, Switzerland
E-mail: raetsch@inf.ethz.ch

## *Better exploration, more security with k-mer analysis of clinical sequencing data and advanced machine learning*

William Ritchie (CNRS).

We recently demonstrated that splitting large sequencing datasets into sequences of length k (a procedure called k-mer decomposition) and exploring these with an adaptive genetic algorithm allowed a more accurate classification and prediction of disease outcome. In addition, this analysis could be performed without a reference genome or transcriptome and was applicable to any sequencing technology. We now demonstrate through advanced machine learning approaches that our k-mer approaches can help explore RNA-seq data more thoroughly, much faster and with much less computational resources than classical bioinformatics approaches. In addition, the use of our approach can help protect patients identity and allows for better exchange of sensitive genomic data.

## *AccuFusion: A new statistical algorithm for detecting gene fusions in RNA-Seq data*

Xiaoping Su (UT MD ANDERSON CANCER CENTER) and Gabriel Malouf (Strasbourg University Hospital and IGBMC).

The advent of high-throughput sequencing (HTS) technologies allows for the detection of gene fusions at unprecedented efficiency.

We developed a highly scalable software, AccuFusion, which implements a robust statistical model-based algorithm for gene fusion discovery by paired-end RNA-Seq. Specifically, a given paired-end read alignment is first quantified in terms of the genomic location (L) of the aligned read pair, the distance (D) between the aligned read pair of the fragment (insert) and the orientation (O) of the read pair. The specific pattern in (L, D, O) space is used as a constraint to define the discordant read pair. This algorithm begins by detecting and clustering discordant read pairs that support the same fusion event (e.g. BCR-ABL1) and selects the discordant read clusters as fusion candidates. Next, a greedy strategy is applied to define the boundaries of discordant read clusters to address the existence of multiple fusion products with different fusion junctions in the same fusion partners. An in-silico sequence generated by using the consensus of reads within discordant read clusters for each fusion candidate is used to detect breakpoint-spanning reads. Those steps and other filtering metrics are used to reduce the false positive fusion candidates.

## *Genomic loci susceptible to systematic sequencing bias in clinical whole genomes*

Timothy Freeman (The University of Sheffield), Dennis Wang (The University of Sheffield) and Jason Harris (Personalis Inc.).

Accurate high-throughput sequencing (HTS) of genetic variants is vital for research and medicine. Certain genomic positions can be prone to higher rates of systematic sequencing and alignment bias, resulting in false-positive variant calls. Current standard practices to identify loci that cannot be sequenced accurately utilize consensus between different sequencing methods as a proxy for sequencing confidence. These practices have significant limitations that remain unaddressed.

We have developed a novel statistical method, summarizing sequenced reads from whole genome clinical samples in “Incremental Databases” maintaining patient confidentiality. Allele statistics were cataloged for each genomic position that consistently showed systematic biases with the corresponding HTS pipeline.

We found systematic biases at ~1-3% of the human autosomal genome across five patient cohorts. We identified how susceptible different genomic regions were to systematic biases, including large homopolymer flanks (odds ratio=23.29-33.69) and the NIST high-confidence genomic regions (odds ratio=0.154-0.191). We confirmed our predictions on gold-standard reference genomes and demonstrated these systematic biases lead to suspect variant calls within clinical panels.

This study implements a novel method to enhance quality control of sequenced samples, by flagging variants displaying systematic sequencing biases for further analysis or exclusion. This work has recently been accepted for publication in Genome Research.

## *Epigenomic enrichment analysis using Bioconductor*

Dario Righelli (Department of Statistics, University of Padua), Ben Johnson (Van Andel Research Institute), Claudia Angelini (Istituto per le applicazioni del Calcolo "M. Picone", National Research Council), Tim Triche (Van Andel Research Institute), Lucia Peixoto (Department of Biomedical Sciences, Elson S. Floyd College of Medicine, Washington State University) and Davide Risso (Department of Statistics, University of Padua).

The low cost of sequencing has made epigenomics a powerful instrument for the study of chromatin accessibility, histone modifications, and other epigenetic mechanisms.

Several tools have been developed to tackle the main steps of the computational and statistical analyses. The ENCODE consortium has developed a set of best practices and experimental guidelines for transcriptomic and epigenomic experiments. However, these guidelines focus on signal discovery and there is still no consensus on the best practices for the detection of differential signal, e.g., when comparing tissues under different conditions or following different treatments.

Here, we focus in particular on ATAC-seq/SONO-seq and Histone Modification ChIP-seq. We present three complete workflows for the differential analysis of epigenomic data, based on the Bioconductor packages DiffBind, DEScan2, and CSAW, respectively. We compare the approaches on real data and provide guidelines on how to best apply each pipeline to one's data. Moreover, we provide some indication on lower-level, and often neglected, steps of the analysis, such as the impact of the aligner, read de-duplication, and normalization on the final results.

## *VariantStore: A Large-Scale Genomic Variant Search Index*

Prashant Pandey (Carnegie Mellon University), Yinjie Gao (Carnegie Mellon University) and Carl Kingsford (Carnegie Mellon University).

The ability to efficiently query genomic variants from thousands of samples is critical to achieving the full potential of many medical and scientific applications such as personalized medicine. Performing variant queries based on positions in the reference or sample sequences is at the core of these applications. Efficiently supporting variant queries across thousands of samples is computationally challenging. Most solutions only support queries based on the reference sequence and the ones that support queries across multiple samples do not scale to data containing more than a few thousand samples. We present VariantStore, a system for efficiently indexing and querying genomic variants and their sequences in either the reference or sample-specific positions. We show the scalability of VariantStore by indexing genomic variants from the TCGA-BRCA project containing 8640 samples and 5M variants in 4 Hrs and the 1000 genomes project containing 2500 samples and 924M variants in 3 Hrs. Querying for variants in a gene takes between 0.002 – 3 seconds using memory only 10% of the size of the full representation.

## *Defective and Intact HIV genome Assembler (DIHIVA), a pipeline to annotate and classify HIV genomes based on NGS data.*

Victor Ramos (The Rockefeller University), Thiago Yukio Kikuchi Oliveira (The Rockefeller University), Christian Gaebler (The Rockefeller University), Pilar Mendoza Daroca (The Rockefeller University) and Michel C. Nussenzweig (The Rockefeller University).

Studies combining PCR techniques and next-generation DNA sequencing (NGS) to characterize the HIV genetic diversity and understand the reservoir dynamics have been extensively reported. In such investigation thousands of sequences are generated, requiring automation in the characterization process. Here we present the Defective and Intact HIV genome Assembler (DIHIVA), a pipeline to annotate and classify HIV genomes based on NGS data. First, we remove PCR amplification and perform error correction using BBtools. After, a quality-control check is carried out by Trim Galore package to trim Illumina adapters and low-quality bases. BBtools is also used to remove possible contaminant reads using HIV genome sequences obtained from Los Alamos HIV database. We then use SPAdes, a k-mer based assembler, to reconstruct the HIV-1 sequences. The longest assembled contig is aligned via BLAST to the Los Alamos HIV genome database, in order to set it in the forward orientation. Finally, the HIV genome is annotated against Hxb2 using BLAST. In the end, sequences are classified as intact or defective, which could be Major Splicing Donor (MSD) Mutation, Non-functional, Missing Internal Genes or Weird. In order to provide access to non-computer specialists we developed a user-friendly interface for the pipeline using R/Shiny.

## *Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis*

Kristoffer Sahlin (Department of Mathematics, Stockholm University), Botond Sipos (Oxford Nanopore Technologies Ltd.), Phillip James (Oxford Nanopore Technologies Ltd.), Daniel Turner (Oxford Nanopore Technologies Ltd.) and Paul Medvedev (The Pennsylvania State University).

Oxford Nanopore (ONT) is a leading long-read technology which has been revolutionizing transcriptome analysis through its capacity to sequence the majority of transcripts from end-to-end. This has greatly increased our ability to study the diversity of transcription mechanisms such as transcription initiation, termination, and alternative splicing. However, ONT still suffers from high error rates which have thus far limited its scope to reference-based analyses. When a reference is not available or is not a viable option due to reference-bias, error correction is a crucial step towards the reconstruction of the sequenced transcripts and downstream sequence analysis of transcripts. In this paper, we present a novel computational method to error-correct ONT cDNA sequencing data, called isONcorrect. IsONcorrect is able to jointly use all isoforms from a gene during error correction, thereby allowing it to correct reads at low sequencing depths. We are able to obtain an accuracy of 98.9-99.6%, demonstrating the feasibility of applying cost-effective cDNA full transcript length sequencing for reference-free transcriptome analysis.

## *Leveraging Hi-C and Whole Genome Shotgun Sequencing for Double Minute Chromosome Discovery*

Matthew Hayes (Xavier University of Louisiana), Angela Nguyen (Xavier University of Louisiana), Ethan Tran (Xavier University of Louisiana), Derrick Mullins (Xavier University of Louisiana) and Chindo Hicks (Louisiana State University Health Sciences Center - New Orleans).

Double minute chromosomes are highly amplified oncogenic, acentric, extrachromosomal DNA that are frequently observed in the cells of numerous cancer types. Algorithmic discovery of double minutes (DM) can potentially improve bench-derived therapies for cancer treatment. A hindrance to this task is that DMs evolve, yielding circular chromatin that shares segments from progenitor double minutes. This creates multiple double minutes in the tumor genome that are distinct, but that share loci for overlapping amplicon coordinates. Existing DM discovery algorithms (that use only whole genome sequencing data) can potentially misclassify DMs that share overlapping coordinates. In this study, we describe a method called HolistIC that predicts double minutes in tumor genomes by integrating whole genome shotgun sequencing (WGS) and Hi-C sequencing data. This resolves ambiguity in double minute prediction that exists when using WGS data alone, a limitation of existing methods for this problem. We implemented and tested our algorithm on the tandem Hi-C and WGS datasets of two cancer datasets and a simulated dataset. Our results show that HolistIC can distinguish between double minutes that share amplicon coordinates, an advance over current methods for this problem.

# *Model-based analysis of sample index hopping reveals its widespread artifacts in multiplexed single-cell RNA-sequencing*

Rick Farouni (Department of Human Genetics, McGill University and Genome Quebec Innovation Centre), Haig Djambazian (Department of Human Genetics, McGill University and Genome Quebec Innovation Centre), Jiannis Ragoussis (Department of Human Genetics, McGill University and Genome Quebec Innovation Centre) and Hamed S. Najafabadi (Department of Human Genetics, McGill University and Genome Quebec Innovation Centre).

We introduce a statistical model for the estimation of sample index-hopping rate in multiplexed droplet-based single-cell RNA sequencing data and probabilistic inference of the true sample of origin of the hopped reads. Across the datasets we analyzed, we estimate the sample index hopping probability to range between 0.003–0.009, a small number that counter-intuitively gives rise to a large fraction of "phantom molecules" – in more than 25% of samples, the fraction of phantom molecules exceeds 8%, with this fraction reaching as high as 85% in low-complexity samples. These phantom molecules lead to widespread complications in downstream analyses, including transcriptome mixing across cells, emergence of phantom copies of cells from other samples, and misclassification of empty droplets as cells. We demonstrate that our model-based approach can correct for these artifacts by accurately purging the majority of phantom molecules from the data. An R package implementing our approach along with reproducible R markdown notebooks of the analyses are available at https://csglab.github.io/PhantomPurgeR/.

## *ntJoin: Fast and lightweight assembly-guided scaffolding using minimizer graphs*

Lauren Coombe (BC Cancer Genome Sciences Centre), Vladimir Nikolić (BC Cancer Genome Sciences Centre), Justin Chu (BC Cancer Genome Sciences Centre), Inanc Birol (BC Cancer Genome Sciences Centre) and Rene Warren (BC Cancer Genome Sciences Centre).

The ability to generate high-quality genome sequences is cornerstone to modern biological research. Even with recent advancements in sequencing technologies, many genome assemblies are still not achieving reference-grade. Recently, we have developed ntJoin, a tool that leverages structural synteny between a draft assembly and reference sequence(s) to contiguate and correct the former with respect to the latter. Instead of alignments, ntJoin uses a lightweight mapping approach based on a graph data structure generated from ordered minimizer sketches. The tool can be used in a variety of different applications, including (but not limited to) improving a draft assembly with a reference-grade genome, a short read assembly with a draft long read assembly, and a draft assembly with an assembly from a closely-related species. When scaffolding a human short read assembly using the reference human genome or a long read assembly, ntJoin improves the NGA50 length 23- and 13-fold, respectively, in under 13 min, using less than 11 GB of RAM. Compared to existing reference-guided scaffolders, ntJoin generates highly contiguous assemblies faster and using less memory. ntJoin is written in C++ and Python, and is freely available at https://github.com/bcgsc/ntjoin.

## *Tumor heterogeneity assessed by sequencing and fluorescence in situ hybridization (FISH) data*

Haoyun Lei (Carnegie Mellon University), Edward Gertz (National Institutes of Health, NCI), Alejandro A. Schaffer (National Institutes of Health, NCI), Xuecong Fu (Carnegie Mellon University), Yifeng Tao (Carnegie Mellon University), Kerstin Heselmeyer-Haddad (National Institutes of Health), Irianna Torres (National Institutes of Health), Xulian Shi (BGI-Shenzhen), Kui Wu (BGI-Shenzhen), Guibo Li (BGI-Shenzhen), Liquin Xu (BGI-Shenzhen), Yong Hou (BGI), Michael Dean (Cancer and Inflammation Program, National Cancer Institute, National Institutes of Health), Thomas Ried (National Institutes of Health) and Russell Schwartz (Carnegie Mellon University).

Computational reconstruction of clonal evolution in cancers has become a crucial tool for understanding how tumors initiate and progress and how this process varies across patients. The field still struggles, however, with special challenges of applying phylogenetic methods to cancers, such as the prevalence and importance of copy number alteration (CNA) and structural variation (SV) events in tumor evolution, which are diffcult to profile accurately by prevailing sequencing methods in such a way that subsequent reconstruction by phylogenetic inference algorithms is accurate. In the present work, we develop computational methods to combine sequencing with multiplex interphase fluorescence in situ hybridization (miFISH) to exploit the complementary advantages of each technology in inferring accurate models of clonal CNA evolution accounting for both focal changes and aneuploidy at whole-genome scales. We demonstrate on simulated data that incorporation of FISH data substantially improves accurate inference of focal CNA and ploidy changes in clonal evolution from deconvolving bulk sequence data. Analysis of real glioblastoma data for which FISH, bulk sequence, and single cell sequence are all available confirms the power of FISH to enhance accurate reconstruction of clonal copy number evolution in conjunction with bulk and optionally single-cell sequence data.

## *A novel mtDNA methylation framework reveals mtDNA methylation in the heavy strand promoter correlating with mitochondrial gene expression*

Romain Guitton (University of Bergen), Gonzalo Nido (University of Bergen) and Charalampos Tzoulis (University of Bergen).

While mitochondrial genetics have been extensively studied, the role of epigenetic regulation of the mitochondrial genome remains largely unexplored. DNA methylation plays a major role in regulating gene expression in the nucleus. The existence and potential role of mitochondrial DNA (mtDNA) methylation remain, however, controversial. Here, we employ a combination of dry and wet lab methodologies to assess mtDNA methylation in human prefrontal cortex from neurological healthy controls (n =27) and individuals with Parkinson's disease (PD) (n = 27). A specific double-mapping approach to analyse whole-genome bisulfite sequencing data efficiently removes false-positive misalignments of nuclear mitochondrial pseudogene sequences (NUMT).The ultra-deep mtDNA coverage (16842±4905) reveals an overall lack of mtDNA methylation with the exception of a single CpG hotspot at position m.545 on the heavy-strand-promoter. Furthermore, the m.545 methylation hotspot is validated using a combination of methylation-sensitive DNA restriction and quantitative PCR. In addition, we show that the methylation levels of m.545 positively correlate with mtDNA gene expression. We detect no differences in m.545 methylation between PD cases and controls. Our findings suggest that, mtDNA methylation occurs in human prefrontal cortex exclusively on the m.545 position of the heavy-strand-promoter and it may play a role in mitochondrial gene expression.

## *The intestinal Th17 population and its role in extra-intestinal autoimmune disease*

Linglin Huang (Department of Biostatistics Harvard T.H. Chan School of Public Health), Alexandra Schnell (Evergrande Centre for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital), Meromit Singer (Department of Data Sciences, Dana-Farber Cancer Institute), Rafael Irzarry (Department of Data Sciences, Dana-Farber Cancer Institute), Aviv Regev (Broad Institute of MIT and Harvard) and Vijay Kuchroo (Evergrande Centre for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital).

Background The Il17-producing T helper subset (Th17) is a well-established driver of multiple autoimmune diseases. At homeostasis most Th17 cells are found in the lamina propria of the intestine. Interestingly, recent studies strongly suggest an involvement of the intestinal Th17 population in extra-intestinal autoimmune diseases. However, the mechanism in which intestinal Th17 cells can drive autoimmune tissue inflammation in peripheral sites remains to be elucidated. Results We performed single-cell RNA sequencing of splenic, lymph node-derived, intestinal and central nervous system-infiltrating Th17 cells at homeostasis and during experimental autoimmune encephalomyelitis (EAE). Accordingly, we built an analysis pipeline for a comprehensive characterization of the Th17 population. In particular, we identified tissue-specific Th17 signatures and characterized the heterogeneity of Th17 cells in each tissue at homeostasis. We also described the changes in transcriptional profiles upon EAE in different tissues. Furthermore, we used single-cell TCR sequencing with cell hashing to characterize Th17 cell within-tissue clonal expansion as well as across-tissue migration at homeostasis and during EAE. Conclusions Our study provides an extensive single-cell survey of tissue Th17 cells during homeostasis and EAE. Th17 cells show transcriptomic heterogeneity within and across tissues, and there are evidences for cell migration to peripheral sites during EAE.

## *Chemical Safety Screening Using High-Throughput Transcriptomics*

Logan J. Everett (U.S. Environmental Protection Agency), Joshua A. Harrill (U.S. Environmental Protection Agency), Derik Haggard (Oak Ridge Institute for Science and Education), Joseph Bundy (U.S. Environmental Protection Agency), Imran Shah (U.S. Environmental Protection Agency), Richard Judson (U.S. Environmental Protection Agency), Thomas Sheffield (U.S. Environmental Protection Agency) and Woodrow Setzer (U.S. Environmental Protection Agency).

High-Throughput Transcriptomics (HTTr) is emerging as a cost-effective method for broadly assessing chemical toxicity across many target pathways and modes of action in a single assay. US EPA has designed an in vitro chemical screening protocol using targeted RNA-seq wherein batches of chemicals are tested in 8-point concentration series in triplicate using a 384-well plate format. Using a pilot set of 42 chemicals exposed to MCF7 cells, we have demonstrated the utility of this platform for predicting both biological target and overall point of departure for each chemical. US EPA has subsequently scaled this screening platform to several thousand chemicals across multiple cell lines. Our recent work on this growing collection of data has included: 1) development of an open-source pipeline for rapid and robust processing of targeted RNA-seq data; 2) comparison of multiple analytical methods for estimating differential expression and dose-response models; 3) quantification of reproducibility across platforms and methods; and 4) signature-level analysis methods to summarize results and link findings to interpretable biology and known hazards. Notably, the results show that aggregating signal at the signature level improves reproducibility and reduces uncertainty in screening results. This abstract does not necessarily reflect US EPA policy.

## *Single-cell copy number lineage tracing enabling gene discovery*

Ken Chen (The University of Texas MD Anderson Cancer Center), Fang Wang (The University of Texas MD Anderson Cancer Center), Qihan Wang (Rice University), Vakul Mohanty (The University of Texas MD Anderson Cancer Center), Shaoheng Liang (The University of Texas MD Anderson Cancer Center), Jinzhuang Dou (The University of Texas MD Anderson Cancer Center), Jincheng Han (The University of Texas MD Anderson Cancer Center), Darlan Minussi (The University of Texas MD Anderson Cancer Center), Ruli Gao (Houston Methodist Research Institute), Li Ding (Washington University School of Medicine) and Nicholas Navin (The University of Texas MD Anderson Cancer Center).

Aneuploidy plays critical roles in genome evolution. Alleles, whose dosages affect the fitness of an ancestor, will have altered frequencies in the descendant populations upon perturbation. Single-cell sequencing enables comprehensive genome-wide copy number profiling of thousands of cells at various evolutionary stage and lineage. That makes it possible to discover dosage effects invisible at tissue level, provided that the cell lineages can be accurately reconstructed. Here, we present a Minimal Event Distance Aneuploidy Lineage Tree (MEDALT) algorithm that infers the evolution history of a cell population based on single-cell copy number (SCCN) profiles. We also present a statistical routine named lineage speciation analysis, which facilitates discovery of fitness-associated alterations and genes from SCCN lineage trees. We assessed our approaches using a variety of single-cell datasets. Overall, MEDALT appeared more accurate than phylogenetics approaches in reconstructing copy number lineage. From the single-cell DNA-sequencing data of 20 triple-negative breast cancer patients, our approaches effectively prioritized genes that are essential for breast cancer cell fitness and are predictive of patient survival, including those implicating convergent evolution. Similar benefits were observed when applying our approaches on single-cell RNA sequencing data obtained from cancer patients. The source code of our study is available at https://github.com/KChen-lab/MEDALT.

## *Optimizing Transcriptomics for High-Throughput Bioactivity Screening of Environmental Chemicals*

Derik Haggard (Oak Ridge Institute for Science and Education, U.S. Environmental Protection Agency), Thomas Sheffield (Oak Ridge Institute for Science and Education, U.S. Environmental Protection Agency), Joshua Harrill (U.S. Environmental Protection Agency), Imran Shah (U.S. Environmental Protection Agency), Richard Judson (U.S. Environmental Protection Agency), Woodrow Setzer (U.S. Environmental Protection Agency) and Logan Everett (U.S. Environmental Protection Agency).

High-throughput transcriptomics (HTTr) shows promise as a useful method in chemical safety assessment due to its broad biological coverage and potential for identifying diverse mechanisms of action for many chemicals. Concentration response profiling also provides information on chemical hazard and potency, which can inform chemical prioritization. US EPA has screened ~2,200 chemicals in 8-point concentration response in MCF-7 cells using the TempO-Seq whole transcriptome assay. Screening was performed in 384-well plate format and included multiple reference samples to assess the biological and technical variability of the platform. We developed several quality control metrics based on mapping rate and read count distributions across probes to flag low-quality samples and estimate that ~94% of samples had sufficient quality. Analysis of reference samples demonstrated high reproducibility across plates and experimental groups, with a median correlation near 0.93. Further refinement of the concentration-response modeling suggests signature-level analyses have predictive power for identifying points of departure for specific mechanisms of bioactivity. This work represents one of the largest transcriptomic chemical safety screens to date, shows promise as a first-tier screening method for hazard identification, and may inform the prioritization of chemicals for additional study. This abstract does not necessarily reflect US EPA policy

## *Inferring copy number substructure from single cell transcriptomics in human tumors*

Ruli Gao (Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, USA 77030), Shanshan Bai (Department of Genitourinary Medical Oncology, UT MD Anderson Cancer Center, Houston TX, USA 77030), Henderson Ying (Department of Head and Neck Surgery, UT MD Anderson Cancer Center, Houston TX, USA 77030), Yiyun Lin (Department of Genetics, UT MD Anderson Cancer Center, Houston TX, USA 77030), Tapsi Kumar (Department of Genetics, UT MD Anderson Cancer Center, Houston TX, USA 77030), Min Hu (Department of Genetics, UT MD Anderson Cancer Center, Houston TX, USA 77030), Emi Sei (Department of Genetics, UT MD Anderson Cancer Center, Houston TX, USA 77030), Alexander Davis (Department of Genetics, UT MD Anderson Cancer Center, Houston TX, USA 77030), Fang Wang (Department of Bioinformatics and Computational Biology, UT MD Anderson Cancer Center, Houston TX, USA 77030), Jennifer Rui Wang (Department of Head and Neck Surgery, UT MD Anderson Cancer Center, Houston TX, USA 77030), Ken Chen (Department of Bioinformatics and Commputational Biology, UT MD Anderson Cancer Center, Houston TX, USA 77030), Stacey Moulder (Department of Breast Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA 77030), Stephen Lai (Department of Head and Neck Surgery, UT MD Anderson Cancer Center, Houston TX, USA 77030) and Nicholas Navin (Department of Genetics, UT MD Anderson Cancer Center, Houston TX, USA 77030).

High-throughput single cell transcriptomics analysis is widely used to study human tumors, however a major challenge is distinguishing the stromal cells from the malignant cancer cells, as well as clonal substructure within tumors. To address this challenge, we developed an integrative Bayesian segmentation approach, COPYKAT to estimate genomic copy numbers at 5MB resolution from high-throughput single cell RNA-seq data. We applied COPYKAT to 39,709 single cells from 16 tumors across 4 cancer types, including premalignant and triple-negative breast cancers, pancreatic ductal adenocarcinomas, and anaplastic thyroid cancer. From these data we could accurately (98%) classify tumor cells from stromal cells. In three TNBC tumors COPYKAT resolved multiple clonal subpopulations of genotypes that differed in expression of breast cancer genes and enrichment of cancer hallmarks including EMT and hypoxia. These data show that COPYKAT can accurately resolve clonal copy number substructure in tumors and classify tumor and normal cells in a variety of human cancers.

## *Benchmarks for Challenging Genome Regions and Structural Variants*

Justin Zook (National Institute of Standards and Technology), Justin Wagner (National Institute of Standards and Technology), Nathan Olson (National Institute of Standards and Technology), Marc Salit (Joint Initiative for Metrology in Biology) and Genome In A Bottle Consortium (National Institute of Standards and Technology).

While Next Generation Sequencing with short reads can achieve high accuracy in most of the human genome, specialized sequencing and/or analysis methods are required for calling structural variants (SVs) and variants in repetitive regions such as tandem repeats, homologous genes, and pseudogenes. The Genome In A Bottle (GIAB) consortium was established to authoritatively characterize genomes for use in method validation and benchmarking. Here, we use linked- and long-reads to establish new benchmarks for structural variants and small variants in challenging genomic regions. The new SV benchmark set includes ~10,000 insertions and deletions >50bp in size inside 2.66 Gbp benchmark regions supported by diploid assemblies. Many of these SVs were sequence-resolved by long reads and de novo assembly. For the new small variant benchmark, long and linked-read data added >300,000 new variants and 189 million bp, over half in regions that are difficult to map with short reads, including coding regions in medically important genes (e.g., PMS2 and CYP21A2). To demonstrate the utility of these benchmark sets, we have compared variant calls from a variety of technologies to the benchmark sets, and we have shown that they reliably identify false positives and false negatives.

## *RResolver: short read repeat resolution with sliding window*

Vladimir Nikolic (Canada's Michael Smith Genome Sciences Centre at BC Cancer), Justin Chu (Canada's Michael Smith Genome Sciences Centre at BC Cancer), Johnathan Wong (Canada's Michael Smith Genome Sciences Centre at BC Cancer), Lauren Coombe (Canada's Michael Smith Genome Sciences Centre at BC Cancer), Ka Ming Nip (Canada's Michael Smith Genome Sciences Centre at BC Cancer), Rene Warren (Canada's Michael Smith Genome Sciences Centre at BC Cancer) and Inanc Birol (Canada's Michael Smith Genome Sciences Centre at BC Cancer).

De novo genome assembly is an important tool for modern molecular biology. It is also a useful method for studying genomes with high variation, such as in the context of cancer, as it is not biased by any reference. De novo short read assemblers commonly use de Bruijn graphs where sequences initially overlap by the number of bases specified with parameter k, followed by the identification of unambiguous walks on the graph. This parameter selection is a trade-off between connectivity and contiguity, but is inherently a crude approach that only works on average. Here, we present RResolver, an algorithm that takes a short read assembly with a fixed k as input, and then uses a k value close to the read size to resolve repeats in the graph. This is accomplished by a sliding window along possible paths to determine which ones are unsupported by reads and should be removed. When used with the popular short read assembler ABySS, RResolver takes 4 hours (6% of the whole pipeline) with 48 threads, and 40GB memory to run on a 150bp read human assembly. It improves contiguity (NGA50) by 5% while reducing misassemblies by 5% in the final result.

## *Comparative analysis of workflow management systems in production genomics research and the clinic*

Azza E Ahmed (University of Khartoum), Joshua Allen (University of Illinois at Urbana-Champaign), Saurabh Baheti (Mayo Clinic), Tajesvi Bhat (University of Illinois at Urbana-Champaign), Matthew A Bockol (Mayo Clinic), Prakruthi Burra (University of Illinois at Urbana-Champaign), Travis M Drucker (Mayo Clinic), Steven N Hart (Mayo Clinic), Jacob R Heldenbrand (University of Illinois at Urbana-Champaign), Matthew Hudson (University of Illinois at Urbana-Champaign), Michael Kalmbach (Mayo Clinic), Gregory Kapraun (Mayo Clinic), Eric W Klee (Mayo Clinic), Katherine Kendig (University of Illinois at Urbana-Champaign), Matthew Kendzior (University of Illinois at Urbana-Champaign), Nathan R Mattson (Mayo Clinic), Christian A Ross (Mayo Clinic), Sami M Sharif (University of Khartoum), Eric D Wieben (Mayo Clinic), Mathieu Wiepert (Mayo Clinic), Ramshankar Venkatakrishnan (University of Illinois at Urbana-Champaign), Faisal M Fadlelmola (Center for Bioinformatics & Systems Biology, Faculty of Science, University of Khartoum) and Liudmila S Mainzer (University of Illinois at Urbana-Champaign).

The changing landscape of genomics research and clinical practice warrants computational pipelines that efficiently orchestrate complex analyses of large data volumes across heterogeneous environments. We performed a systematic evaluation of the popular bioinformatics Workflow Management Systems (WfMS): Nextflow, CWL and WDL and some of their executors, along with Swift/T, a WfMS commonly used in high-scale physics applications. A genomic variant calling pipeline and a scalability-testing framework were developed in each and run locally, on an HPC cluster and in the cloud. This allowed contrasting the four in terms of language expressiveness, modularity, scalability, robustness, reproducibility, interoperability, ease of development, and adoptability. This evaluation helps guide the choice of the appropriate WfMs as driven by the interplay of the workflow language, executor and computational ecosystem, as well as the collaborative practices of our field. We also arrived at a set of design considerations to further improve WfMS deployment in research and clinical settings. As the community and its needs continue to evolve along with computational infrastructure, WfMS will evolve, especially those with permissive licenses. Just as the dataflow paradigm and containerization are now widely used, this field will continue to see innovations elsewhere, such as big data technologies and portability.

## *Comprehensive and streamlined quality control for single-cell RNA data*

Shruthi Bandyadka (Boston University, Bioinformatics Graduate Program), Yusuke Koga (Boston University, Bioinformatics Graduate Program), Anastasia Leshchyk (Boston University, Bioinformatics Graduate Program), Rui Hong (Boston University School of Medicine), Zhe Wang (Boston University School of Medicine) and Joshua Campbell (Boston University School of Medicine).

Performing comprehensive quality control is necessary to remove poor quality cells in single-cell RNA sequencing (scRNA-seq) data. Artifacts in the scRNA-seq data, such as doublets or ambient RNA, can also hinder downstream clustering and marker selection and need to be assessed. While several algorithms have been developed to perform various quality control tasks, they are only available in different packages across various programming environments. We have built an easy-to-use pipeline in the singleCellTK package that generates a comprehensive summary of quality control metrics from several existing tools. Our pipeline is able to import data from various preprocessing tools, compute general quality control metrics, detect doublets, and predict and correct for ambient RNA contamination for each cell. Further, to make the data accessible in down-stream analysis workflows, it exports the results to common data structures including the SingleCellExperiment object for analysis in R, the AnnData object for analysis in Python, as well as flat text files for use in any generic workflow.  Overall, this pipeline will help streamline and standardize quality control analyses for scRNA-seq data.

## *PopDel detects large deletions jointly in tens of thousands of genomes*

Sebastian Niehus (Berlin Institute of Health (BIH) / Charité – Universitätsmedizin Berlin), Janina Schoenberger (Berlin Institute of Health (BIH) / Charité – Universitätsmedizin Berlin), Hákon Jónsson (deCODE genetics/Amgen Inc), Eythór Björnsson (deCODE genetics/Amgen Inc), Doruk Beyter (deCODE genetics/Amgen Inc), Hannes P. Eggertsson (deCODE genetics/Amgen Inc), Patrick Sulem (deCODE genetics/Amgen Inc), Kári Stefánsson (deCODE genetics/Amgen Inc), Bjarni V. Halldórsson (deCODE genetics/Amgen Inc) and Birte Kehr (Berlin Institute of Health (BIH) / Charité – Universitätsmedizin Berlin).

Catalogs of genetic variation for large numbers of individuals are a foundation for modern research on human diversity and disease. Creating such catalogs for small variants from whole-genome sequencing (WGS) data is now commonly done for thousands of individuals collectively. We have transferred this joint calling idea from SNPs and indels to larger deletions and developed the first joint calling tool, PopDel, that can detect and genotype deletions in WGS data of tens of thousands of individuals simultaneously as demonstrated by our evaluation on data of up to 49,962 human genomes. Good sensitivity, precision and the correctness of genotypes are demonstrated by extensive tests on simulated and real data and comparison to other state-of-the-art SV-callers. PopDel detects deletions in HG002 and NA12878 with high sensitivity while maintaining a low false positive rate as shown by our comparison with different high-confidence reference sets. On data of up to 6,794 trios, inheritance patterns are in concordance with Mendelian inheritance rules and exhibit a close to ideal transmission rate. PopDel reliably reports common, rare and de novo deletions. Therefore, PopDel enables routine scans for deletions in large-scale sequencing studies and we are currently in the process of implementing the detection of other SV-types.

## *Metalign: Efficient alignment-based metagenomic profiling via containment min hash*

Nathan Lapierre (University of California, Los Angeles), Mohammed Alser (ETH Zurich), Eleazar Eskin (University of California, Los Angeles), David Koslicki (Pennsylvania State University) and Serghei Mangul (University of Southern California).

Whole-genome shotgun sequencing enables the analysis of microbial communities in unprecedented detail, with important implications in medicine and ecology. Predicting the presence and relative abundances of microbes in a sample, known as “metagenomic profiling”, is a critical step in microbiome analysis. Existing profiling methods have been shown to suffer from poor false positive or false negative rates, while alignment-based approaches are often considered accurate but computationally infeasible. Here we present a novel method, Metalign, that addresses these concerns by performing efficient alignment-based metagenomic profiling. Metalign employs a high-speed, high-recall pre-filtering method based on the mathematical concept of Containment Min Hash to reduce the reference database size dramatically before alignment, followed by a method to estimate organism relative abundances in the sample by handling reads aligned to multiple genomes. We show that Metalign achieves significantly improved results over existing methods on simulated datasets (Figure 1) from a large benchmarking study, CAMI, and performs well on in vitro mock community data and environmental data from the Tara Oceans project. Metalign is freely available at https://github.com/nlapier2/Metalign, and via bioconda.

## *Variant calling and assembly of the SARS-CoV-2 genomes using highly accurate long reads*

Elizabeth Tseng (Pacific Biosciences), Armin Toepfer (Pacific Biosciences), Michael Brown (Pacific Biosciences), Zev Kronenberg (Pacific Biosciences), Ivan Sovic (Pacific Biosciences), William Rowell (Pacific Biosciences), John Harting (Pacific Biosciences), Joan Wilson (Pacific Biosciences), Ting Hon (Pacific Biosciences), Janet Ziegle (Pacific Biosciences), Primo Baybayan (Pacific Biosciences), Kevin Eng (Pacific Biosciences), Lei Zhu (Pacific Biosciences) and Jason Underwood (Pacific Biosciences).

SARS-CoV-2 is a single-stranded, 30 kb RNA virus from the coronavirus family that is currently causing a worldwide pandemic, with over 3 million confirmed cases and 212k deaths as of April 2020. Complete and accurate sequencing of the viral genome allows epidemiological tracing and discovery of mutations that may be important for antiviral and vaccine research.

PacBio's SMRT Sequencing uses circular consensus sequencing that can generate long, highly accurate (HiFi) reads for SARS-Cov-2 genomes. We generated several in-house datasets using amplicon-based approaches with insert sizes ranging from 400 bp to several kb long, as well as hybridization-based approaches that can capture fragments of 5 kb or longer.

We tested several variant calling methods including PacBio's own MinorVariant (originally developed for HIV quasi-species detection) and LoFreq. We show that regardless of amplicon size, reference-guided mapping approaches can call SNPs accurately at as low as 10-fold coverage under the assumption of a single, consensus genome. Higher coverages are required for identifying minor variants spiked in at 1%, 10%, and 50% frequency. Further, we compare the results on existing assemblers and show that even coverage and longer sequences are desired for non-guided, de novo assembly.

## *Comprehensive benchmarking of de novo assembly methods for eukaryotic genomes*

Dean Southwood (Macquarie University), Siu Fai Lee (Macquarie University; CSIRO), Rahul Rane (CSIRO; University of Melbourne; Macquarie University), John Oakeshott (CSIRO) and Shoba Ranganathan (Macquarie University).

Assembling reference-quality, chromosome-level genomes for both model and novel eukaryotic organisms is an increasingly achievable task for single research teams. However, the broad variety of sequencing technologies, assembly algorithms, and post-assembly processing tools currently available means that there is no clear consensus on a best-practice computational protocol for eukaryotic de novo genome assembly. Here, we provide a comprehensive benchmark of 28 state-of-the-art assembly and polishing packages, in various combinations, when assembling four eukaryotic genomes using both next generation (Illumina HiSeq) and third generation (Oxford Nanopore Technologies R9 and PacBio RSII) sequencing data. Recommendations are made for the most effective tools given particular genome constraints, such as repeat content, GC%, sequencing depth, and genome size, and against common assessment metrics such as contiguity, computational performance, gene completeness, and reference reconstruction. We also present a Snakemake-based pipeline for eukaryotic genome assembly, pyro, to further assist future de novo assembly and package benchmarking projects.

## *Reference-guided transcript discovery and quantification for long read RNA-Seq data*

Ying Chen (Genome Institute of Singapore), Ploy Pratanwanich (Genome Institute of Singapore), Fei Yao (Genome Institute of Singapore), Yuk Kei Wan (Genome Institute of Singapore), Hwee Meng Low (Genome Institute of Singapore), Viktoriia Iakovleva (Genome Institute of Singapore), Lixia Xin (Duke NUS Medical School), Puay Leng Lee (Genome Institute of Singapore), Qiang Yu (Genome Institute of Singapore), Torsten Wüstefeld (Genome Institute of Singapore), Wee Siong Sho Goh (Genome Institute of Singapore), Boon Hsi Sarah Ng (Genome Institute of Singapore) and Jonathan Göke (Genome Institute of Singapore).

Transcriptome profiling is one of the most frequently used technologies and key to interpreting the function of the genome in human diseases. However, quantification of transcript expression with short read RNA-sequencing remains challenging as different transcripts from the same gene are often highly similar. Nanopore RNA Sequencing reduces the complexity of transcriptome profiling with ultra-long reads that can cover the full length of the isoforms. The technology has a high sequencing error rate and often generates shorter, fragmented reads due to RNA degradation, however, currently no specific transcript quantification method exists for such data. Here, we present bambu, a long read isoform discovery and quantification method. Bambu performs probabilistic assignment of reads to annotated and novel transcripts across samples to improve the accuracy of transcript expression estimates. We apply our method to cancer cell line data with spike-in controls, and compare the results with estimates obtained from short read data. Bambu recovered annotated isoforms from spike-ins and showed consistency in gene expression estimation with existing methods for short read RNA-Sequencing data, but improved accuracy in transcript expression estimation. The method is implemented in R (https://github.com/GoekeLab/bambu), enabling simple, fast, and accurate analysis of long read transcriptome profiling data.

## *distinct: a method for differential analyses via hierarchical permutation tests, with applications to single-cell data*

Simone Tiberi (University of Zurich), Helena L Crowell (University of Zurich), Pantelis Samartsidis (University of Cambridge), Lukas M Weber (Johns Hopkins Bloomberg School of Public Health) and Mark Robinson (University of Zurich).

We present distinct, a statistical method to perform, via hierarchical permutation tests, differential analyses between groups of densities. distinct is a general and flexible tool: due to its fully non-parametric nature, which makes no assumptions on how the data was generated, it can be applied to a variety of datasets. It is particularly suitable to perform differential analyses on single cell data, such as single cell RNA sequencing (scRNA-seq) and high-dimensional flow or mass cytometry (HDCyto) data. While most methods for differential expression target differences in the mean abundance between conditions, single-cell data can show more complex variations. distinct, by comparing full distributions, identifies, both, differential patterns involving changes in the mean, as well as more subtle variations that do not involve the mean (e.g., unimodal vs. bi-modal distributions with the same mean). distinct explicitly models the variability between samples (i.e., biological replicates), can adjust for covariates (e.g., batch effects) and allows multi-group (i.e., >2 groups) comparisons. We will present results, based on scRNA-seq and HDCyto simulated and experimental datasets, where distinct outperforms several competitors and is able to detect more patterns of differential expression compared to canonical differential methods. distinct is freely available as a Bioconductor R package.

## *Improving the efficiency of de Bruijn graph construction using compact universal hitting sets*

Yael Ben Ari (Tel Aviv University), Yaron Orenstein (Ben Gurion University of the Negev), Lianrong Pu (Tel Aviv University) and Ron Shamir (Tel Aviv University).

Very large deep sequencing datasets have become ubiquitous in biomedical research, and efficient algorithms are continuously developed to handle them. Many of them use minimizers to obtain speed-up and save memory usage. Recently, we suggested the use of universal hitting sets (UHS) to potentially improve over minimizers. Here, we showcase the benefit of UHS in the de Bruijn graph construction step of genome assembly by the MSP algorithm. We show that using UHS instead of minimizers leads to faster runtime, more balanced bin partitions and less overall memory.

## *FastqCLS: Fastq Compressor by reordering reads to increase compression ratio for Long-read Sequencing data*

Dohyeon Lee (Pusan National University) and Giltae Song (Pusan National University).

A vast amount of genome sequencing data has been produced over the past decades. This genomic data has exploded with the advent of single-cell analysis. Especially long-read sequencing becomes dominant in genomics, the genomic data explosion requires an enormous level of storage capacity. It costs a lot of time and resources to store and transfer the data and causes a bottleneck in genome sequencing analysis. To resolve this issue, various compression techniques have been proposed to reduce the size of original raw sequencing data in FASTQ, but they have focused on short-read sequencing only. In this study, we design a compression algorithm based on a read reordering using a scoring model to increase the compression ratio for the long-read sequencing data. We evaluate our FastqCLS tool using benchmark datasets that have been commonly used in other FASTQ compression studies. We also include new long-read sequencing data in this validation. We compare our method with existing major FASTQ compression tools. As a result, our FastqCLS outperforms in terms of a compression ratio for storing long-read sequencing data supporting a variety of the read-lengths. FastqCLS can be downloaded from https://github.com/krlucete/FastqCLS.

## *SigRepo: A Structured Platform for the Storage and Streamlined Analysis of Multi-Omics Signatures*

Callen Bragdon (Boston University), Mengze Li (Boston University), Rebekah Miller (Boston University), Stefano Monti (Boston University) and Gary Benson (Boston University).

Biomedical studies routinely generate 'omics' data representing marker profiles (genes, proteins, metabolites, etc.) of the experimental conditions under investigation. The corresponding raw data are usually deposited in public repositories, such as the NIH's Gene Expression Omnibus (GEO) or EMBL-EBI's ArrayExpress. However, extraction of the differential signatures associated with a phenotype often requires significant processing. Furthermore, not all studies make the raw data available and only publish finite sets of marker identifiers. A structured, efficient, and openly accessible platform for signature storage and analysis would thus address an unmet need. We have developed SigRepo, an R-based software system for the representation, storage, and interactive retrieval of omics signatures. The system includes the definition of R6 objects for the representation of signatures and signature collections; a MySQL Database for efficient storage, search, and retrieval of signatures; a suite of R command line functions; and an R Shiny Interface for interactive query and analysis of stored signatures, including signature comparison and pathway enrichment analysis. The system was developed as a Bioconductor package and an associated database, for installation on most architectures, to support both interactive and batch signature-based analyses.

## *Can full-length transcript characterization reveal molecular mechanisms of selection in germinal centre B cells?*

Ozge Gizlenci (University of Cambridge), Louise Matheson (Babraham Institute), Simon Andrews (Babraham Institute), Elisa Monzon-Casanova (Babraham Institute) and Martin Turner (Babraham Institute).

c-Myc is critical for the positive selection of germinal centre (GC) B cells. It indirectly regulates the alternative splicing (AS) of transcripts (e.g. Pkm) via induction of the RNA-binding protein PTBP1. There are still unrevealed alternative isoforms of most transcripts which might mediate functional roles in GC B cells. The changes in the AS in the positively selected GC B cells have been previously addressed using the short-read Illumina sequencing. However, the majority of the transcript variants generated by AS and alternative polyadenylation events were not detected. The long-read sequencing platforms emerged out of the need for full-length sequencing to resolve the complexity of isoforms. In our study, we adapted the long-read sequencing, Oxford Nanopore Technology (ONT), using a Smart-seq2 approach to understand the post-transcriptional regulation in positively selected GC B cells. Using 1ng total RNA, we reached a depth of approximately 2-million reads per sample and detected transcripts from over 9500 genes with at least 5 supporting reads. Our findings support that Smart-seq2 adapted ONT RNA sequencing is a powerful workflow for the identification and quantification of complex isoforms in positively selected GC B cells and for the development of a single-cell long-read sequencing platform for rare cell populations.

## *Fast and scalable RNA-seq splicing analysis for the clinical setting*

Joseph Aicher (University of Pennsylvania), Elizabeth J. Bhoj (Children's Hospital of Philadelphia) and Yoseph Barash (University of Pennsylvania).

Exome sequencing is the most advanced standard-of-care genetic test for people with suspected Mendelian disorders. Yet, the diagnostic rate of exome sequencing is only 31%. A significant challenge with exome sequencing is the difficulty of identifying variants disrupting gene function at the RNA transcript level, especially for RNA splicing. RNA-seq is a promising molecular test for increasing the diagnostic rate by directly measuring changes in RNA splicing that could cause disease. This raises the challenge of developing a statistically principled approach to clinical RNA-seq splicing analysis that is also fast and scalable to allow for comparison to RNA-seq controls from large-scale genomic studies such as GTEx v8.

We present MAJIQ-CLIN, a MAJIQ-based toolkit for splicing detection, quantification, and visualization for the clinical setting. MAJIQ captures and quantifies complex (involving more than two splice junctions) and de novo (unannotated) variations missed by most tools. MAJIQ-CLIN provides a principled approach for identifying and prioritizing splicing outliers in patient samples. We describe new incremental, distributed algorithms implemented in these tools that enable analysis of patient samples to tens of thousands of controls in the order of minutes. We demonstrate in simulation studies, population datasets, and patient RNA-seq how MAJIQ-CLIN outperforms previously described approaches.

## *Imputing Optimal Transport Barycenters of Patient Manifolds*

Alexander Tong (Yale University) and Smita Krishnaswamy (Yale University).

Single-cell data is now being collected across many patients in varying conditions. However, data is still relatively expensive. This opens up the opportunity for computational methods to decrease overall cost by inferring a single-cell measurement based on similarity to the meta-data of other similar samples. We examine this problem with an optimal transport perspective. This allows us to leverage a variant of the Sinkhorn algorithm for extremely computationally efficient approximations of transport along discrete manifolds. Our method first constructs the manifold between samples, then aligns this to the manifold of patients, and finally applies this to interpolate a barycenter sample along this manifold. We show first that we are able to better interpolate samples between timepoints than existing methods e.g. Waddington-OT (Schiebinger et al. 2019 Cell) by accounting for structure between multiple timepoints instead of pairs. We then show when the relationship between patients is an inferred manifold, how to impute a patient's single-cell measurements based on other similar single-cell samples by aligning the manifold of patients with that of single-cell measurements. When the manifold of patients exhibits non-linear but intrinsically low-dimensional structure, we are able to more accurately infer a single-cell measurement.

## *Fast Sequence Search Based on Tensor Sketching*

Amir Joudaki (ETH Zurich), Andre Kahles (ETH Zurich) and Gunnar Ratsch (ETH Zurich).

Despite decades of research, the complexity of biological variability and the magnitude of bioinformatics sequence datasets have been an impediment to finding practical algorithms for searching a large corpus of biological sequences with guaranteed accuracy. Our main contributions are a randomized embedding scheme, which allows us to estimate edit distance, and an efficient random linear sketching, which leads to efficient search and clustering algorithms on sequence datasets. The algorithm is simple to implement and outperforms existing hash-based methods for sequence searching tasks.

We represent sequences as tensors over their sub-sequences, use the tensor product to linearly sketch the tensors, and finally implicitly compute the sketch. The subsequences, which are not necessarily contiguous pieces of the sequence, that allows us to derive a worst-case accuracy. The advantages of our tensor-sketching technique are: 1) sketch footprints are an order of magnitude smaller than hash-based 2) tensor sketches can be computed in a streaming fashion, and 3) our framework is more versatile, namely, it is straightforward to introduce arbitrary mismatch penalties. Our results can be viewed as evidence that the proposed tensor sketching framework can be adapted to sequence search in bioinformatics, opening up a wide range of applications.

## *Physlr: Chromosome-level Genome Assemblies Enabled by the Physical Map of Linked Reads*

Amirhossein Afshinfard (BC Cancer Genome Sciences Centre.), Johnathan Wong (BC Cancer Genome Sciences Centre.), Shaun D. Jackman (10x Genomics), Lauren Coombe (BC Cancer Genome Sciences Centre.), Justin Chu (BC Cancer Genome Sciences Centre.), Vladimir Nikolic (BC Cancer Genome Sciences Centre.), Gokce Dilek (University of British Columbia.), Yaman Malkoç (University of British Columbia.), Rene L. Warren (BC Cancer Genome Sciences Centre.) and Inanc Birol (BC Cancer Genome Sciences Centre.).

Map-based (aka hierarchical shotgun) sequencing and assembly approaches from the Sanger sequencing era tend to generate highly contiguous assemblies, essential for a broad range of downstream analysis. However, these methods are expensive, labor-intensive, and time-consuming compared to methods based exclusively on newer high-throughput sequencing (HiTSeq) technologies. Herein, we transfer the concept of physical maps from the earlier map-based assembly approaches into the domain of next-generation sequencing data. We introduce Physlr, a tool that utilizes long-range information provided by linked-read technologies to construct a chromosome-scale physical map de novo, which can then be used to scaffold a draft assembly of any sequencing data. We benchmark Physlr using stLFR linked reads of three human individuals (NA12878, NA24695, and NA24143). For all datasets, Physlr improves across all studied metrics compared to state-of-the-art linked-read scaffolders, and yields chromosome-level contiguity. For NA12878, Physlr results in 56 Mb NG50 (8.3- and 1.19-fold increase over baseline and best comparator (ARCS), respectively), 7.32 Mb NGA50 (0.78- and 0.08-fold increase, respectively), and fewer misassemblies (-6.5%). We also apply Physlr on a long-read Shasta assembly scaffolded with ARCS and find the NG50 and NGA50 increase by 187% (95.5 Mb) and 10% (20.7 Mb), and misassemblies decrease by 28%.

## *Long read sequencing depth analysis*

Rocio Amorin de Hegedus (University of Florida) and Ana Conesa (University of Florida).

With the advent of long read sequencing (LRS) platforms, there has been an increase in precision and a higher throughput sequencing of transcripts. Software SQANTI has been previously developed by the Conesa lab to qualitatively analyze LRS data. However, a quantifying method relying solely on LRS has not been developed. Such method requires a deeper understanding of the underlying distribution of LRS. Our goal is to evaluate how sequencing depth affects transcript distribution, as a first step in developing an LRS methodology for transcript quantification. To assess this, saturation function in NOIseq package was re-designed to evaluate LRS and applied to the Iso-seq Melanoma and Alzheimer (Sequel2) datasets published by PacBio. A total of 9K detected transcripts ( >10 counts) were found in Melanoma, from 700K reads. While 24K detected transcripts (>10 counts) were found in Alzheimer, from a total of 2M reads. Additionally, data showed transcript length quantification bias, with longer transcripts having lower number of reads. Finally, saturation is achieved with Sequel2 levels of coverage (Alzheimer), but not in those generated by previous technologies (Melanoma). This indicates that at Sequel2 level sequencing depths, a significant fraction of the transcriptome can be detected with a significant read coverage.

## *nDSPA: An R/Bioconductor package for quality metrics, preprocessing, visualization, and differential testing analysis of spatial omics data*

Rajesh Acharya (University of Pittsburgh Medical Center Hillman Cancer Center), Tullia Bruno (University of Pittsburgh Department of Immunology) and Riyue Bao (University of Pittsburgh Medical Center Hillman Cancer Center & University of Pittsburgh Department of Medicine).

Immunotherapy has emerged as one of the main treatments for patients with cancer. Understanding the complexity of tumor microenvironment will enable the discovery of new mechanisms and biomarkers. Studies have used spatial profiling technologies such as NanoString Digital Spatial Profiler (DSP) to discover predictors of therapy response. However, lack of analysis tools and proper statistical approaches hinders the effective transformation of spatial omics data into biologically meaningful results. Here, we developed an open-source R/Bioconductor package, nDSPA, specially designed to leverage the spatial elements for data analysis and hypothesis testing. It implements QC, data preprocessing, visualization, and comparison of transcript or protein expression levels from defined regions of interest (ROIs). By evaluating the performance of differential expression identification using various statistical models, we recommend linear mixed effect models incorporating multiple ROIs, the distance between ROIs, and technical batches. nDSPA will be released on GitHub and accompanied by an R Shiny application for interactive data exploration. While built on DSP data, our statistical algorithm can be generalized to other platforms with spatial compartments. We hope our method will provide new insights into the rapidly developing landscape of data-driven biomarker discovery and enable the broad application of spatial omics technologies in translational research.

## *Hi-C Analysis in practice: A workflow to enhance our understanding of the 3D organization of the genome.*

Katharina Hayer (The Children's Hospital of Philadelphia), Brittney Allyn (University of Pennsylvania), Eugene Oltz (Ohio State University), Ahmet Sacan (Drexel University) and Craig Bassing (University of Pennsylvania).

Motivation: Although chromosome conformation capture techniques have become more standardized over the recent years, the computational analyses downstream are still hard to navigate. While some workflows have been widely adopted, it is clear that each of these algorithms only answers a subset of questions one can ask of this kind of data. Therefore, we think it is crucial to leverage several of these approaches to gain a more complete picture.

Methods: Using the Snakemake workflow engine, we are combining the strengths of packages such as the Juicer pipeline, HiCExplorer tools and HOMER. This workflow is modular and can be easily adjusted for different input data or by adding new algorithms. With this workflow, we were able to explore how different read depths impact the detection of chromosome loops at higher resolutions. Additionally, we benchmarked how the number of replicates effect the reproducibility of the results.

Conclusion: We created a workflow that combines reproducibility, quality control, ease of use and comprehensive visualization to address aspects important to any next generation sequencing analysis. We show best practices using publicly available data to help guide researchers when planning their experiments.

## *Single cell transcriptomic profiling identifies molecular phenotypes of newborn human lung cells*

Soumyaroop Bhattacharya (University of Rochester), Jacquelyn Myers (University of Rochester), Cameron Baker (University of Rochester), Mynzhe Guo (Cincinnati Children's Hospital Medical Center), Soula Danopoulos (Children's Hospital Los Angeles), Jason Myers (University of Rochester), Gautam Bandyopadhyay (University of Rochester), Ravi Misra (University of Rochester), Yan Xu (Cincinnati Children's Hospital Medical Center), Steven Romas (University of Rochester Medical Center), Denise Alalam (Children's Hospital Los Angeles), Jeffrey Whitsett (Cincinnati Children's Hospital Medical Center), Gloria Pryhuber (University of Rochester) and Thomas Mariani (University of Rochester).

While animal model studies have extensively defined mechanisms controlling cell diversity in the developing mammalian lung, the limited data available from late stage human lung development represents a significant knowledge gap. Single cell RNAseq generated transcriptional profiles of 5500 cells obtained from two newborn human lungs. Previously frozen single cell isolates were captured, and library preparation was completed on the Chromium 10X system. Data analysis was performed in Seurat, while cell type annotation was performed using the ToppGene. Single cell sequence data from 32000 mouse lung cells were used for comparison to the human data. Transcriptional interrogation of newborn human lung cells readily identified distinct clusters including multiple populations of endothelial, mesenchymal and immune cells. Signature genes from each of these populations were identified. Computational integration of newborn human and postnatal mouse lung development cellular transcriptomes indicated they are highly comparable and facilitated the identification of distinct cellular subtypes. Comparison of the newborn human and mouse cellular transcriptomes also demonstrated cell type-specific differences in maturation states of newborn human lung cells. Matrix fibroblasts could be separated into younger or older cells. Cells with each molecular profile were spatially resolved within newborn human lung tissue.

## *Clinical Variant Detection and Phasing through Automated Long Read Clustering*

Qian Zeng (LabCorp), Lax Iyer (LabCorp), Ingrid Chen (LabCorp), Stan Letovsky (LabCorp), Brian Krueger (LabCorp), John Harting (Pacific Biosciences) and Kristina Weber (Pacific Biosciences).

Clinical Variant Detection and Phasing through Automated Long Read Clustering Qian Zeng*, Lax Iyer*, Ingrid Chen*, Stan Letovsky & Brian Krueger, Laboratory Corporation of America Holdings John Harting & Kristina Weber, Pacific Biosciences
*Equal Contribution

## *xengsort: Fast lightweight accurate xenograft sorting*

Jens Zentgraf (TU Dortmund University) and Sven Rahmann (University of Duisburg-Essen).

With an increasing number of patient-derived xenograft (PDX) models being created and subsequently sequenced to study tumor heterogeneity and to guide therapy decisions, there is a need for methods to separate reads originating from the graft (human) tumor and reads originating from the host species' (mouse) surrounding tissue. Two kinds of methods are in use: On the one hand, alignment-based tools like XenofilteR require that reads are mapped and aligned (by an external mapper/aligner) to the host and graft genomes separately first; the tool itself then processes the resulting alignments and quality metrics (typically BAM files) to assign each read or read pair. On the other hand, alignment-free tools like xenome work directly on the raw read data (typically FASTQ files). Recent studiescompare different approaches and tools, with partially conflicting results.

We here argue that a carefully engineered alignment-free approach using three-way bucketed Cuckoo hashing offers high read classification accuracy, fast running times and a reasonable memory footprint. Software (``xengsort'') is available (MIT license) at http://gitlab.com/genomeinformatics/xengsort.

## *META^2: Memory-efficient taxonomic classification and abundance estimation for metagenomics with deep learning*

Andreas Georgiou (ETH Zürich, Department of Computer Science), Vincent Fortuin (ETH Zürich, Department of Computer Science), Harun Mustafa (ETH Zürich, Department of Computer Science) and Gunnar Rätsch (ETH Zürich, Department of Computer Science).

Taxonomic classification is an important step in the analysis of samples found in metagenomic studies. Conventional mapping-based methods trade off between high memory and low recall, with recent deep learning methods suffering from very large model sizes. We aim to develop a more memory-efficient technique for taxonomic classification. A task of particular interest is abundance estimation. Current methods initially classify reads independently and are agnostic to the co-occurence patterns between taxa. In this work, we also attempt to take these patterns into account. We develop a novel memory-efficient read classification technique, combining deep learning and locality-sensitive hashing. We show that this approach outperforms conventional mapping-based and other deep learning methods for taxonomic classification when restricting all methods to a fixed memory footprint. Moreover, we formulate the task of abundance estimation as a Multiple Instance Learning problem and we extend current deep learning architectures with two types of permutation-invariant MIL pooling layers: a) deepsets and b) attention-based pooling. We illustrate that our architectures can exploit the co-occurrence of species in metagenomic read sets and outperform the single-read architectures in predicting the distribution over taxa at higher taxonomic ranks.

## *Robust and efficient software platform for de novo genomic diversity analysis*

Andrea Parra (Universidad de los Andes), Jorge Duitama (Universidad de los Andes) and Paula Reyes (Agrosavia).

The development of high-throughput sequencing (HTS) technologies has created a supply of large volumes of sequence data. Genotyping by sequencing (GBS) in particular, is a cost-effective strategy to obtain a large number of samples through reduced representation. To use this data, traditionally, variant calling is performed against a reference genome. In the absence of such a reference, which is the case for most species, variant detection must be done de-novo. Several software solutions have been developed for this purpose, but unfortunately most are either inefficient or discover low numbers of polymorphisms. Here we present a new algorithm for efficient analysis of GBS reads without a reference genome. Validation was conducted using publicly available data of different species and different sequencing protocols as well as different population structures. GBS data of an interspecific rice (Oryza sativa) biparental population was analysed. The 10,000 SNVs discovered with an F1-score up to 5x better than other available technologies. DDrad data from Sea Bass was used to successfully validate diversity analysis, by replicating phylogenetic and population structures. This was done at a fraction of the time: processing 1 Mb of data in about ~3.3s. This represents half the time of the next fastest technology.

## *Long-TUC-seq is a robust method for quantification of metabolically labeled full-length isoforms*

Sorena Rahmanian (UC Irvine), Gabriela Balderrama-Gutierrez (UC Irvine), Dana Wyman (UC Irvine), Cassendra McGill (UC Irvine), Kim Nguyen (UC Irvine), Robert Spitale (UC Irvine) and Ali Mortazavi (University of California, Irvine).

The steady state expression of each gene is the result of a dynamic transcription and degradation of that gene. While regular RNA-seq methods only measure steady state expression levels, RNA-seq of metabolically labeled RNA identifies transcripts that were transcribed during the window of metabolic labeling. Whereas short-read RNA sequencing can identify metabolically labeled RNA at the gene level, long-read sequencing provides much better resolution of isoform-level transcription. Here we combine thiouridine-to-cytosine conversion (TUC) with PacBio long-read sequencing to study the dynamics of mRNA transcription in the GM12878 cell line. We show that using long-TUC-seq, we can detect metabolically labeled mRNA of distinct isoforms more reliably than using short reads. Long-TUC-seq holds the promise of capturing isoform dynamics robustly and without the need for enrichment.

## *Single cell integrated analysis of peripheral immune cell populations in the context of aging*

Tanya Karagiannis (Boston University), Stefano Monti (Boston University) and Paola Sebastiani (Boston University).

Disability, age-related diseases, and morbidity are common risk factors with advancing age. Recent studies have shown that age-related disability and diseases are delayed in people living to 100 and especially compressed in those surviving greater than 100 years. Changes in the immune system with age have also been demonstrated, including differences in proportion of immune cell populations and gene expression differences over age. In order to further explore immune cell populations during age progression, we investigated cell subpopulations of the peripheral blood immune system across age using publicly available single cell RNA-sequencing datasets of peripheral blood mononuclear cells from multiple subjects of extreme longevity and younger control ages ranging from 20-80 years. We performed integrated analyses of these datasets to gain higher resolution of subpopulations of immune cells across age and to generate corresponding gene expression signatures for extreme longevity subjects compared to controls. Early findings suggest changes in cell type composition are associated with age even in smaller subpopulations of cells.

## *Improving RNA-seq mapping and haplotype-specific transcript inference using variation graphs*

Jonas A. Sibbesen (University of California Santa Cruz), Jordan Eizenga (University of California Santa Cruz) and Benedict Paten (University of California Santa Cruz).

Current methods for analyzing RNA-seq data are generally based on first mapping the reads to a reference genome or a known set of reference transcripts. However, this approach can bias read mappings toward the reference, which negatively affects downstream analyses such as haplotype-specific expression quantification. One way to mitigate this reference bias is to use variation graphs, which contain both the primary reference and known genetic variants. For RNA-seq data specifically, variation graphs can also be augmented with splice junctions and haplotype-specific transcripts can be embedded as paths. In this work, we introduce a pipeline based on the variation graph (vg) toolkit for both mapping RNA-seq data to spliced variation graphs and inferring the expression of known haplotype-specific transcripts from the mapped reads. We demonstrate that spliced variation graphs reduce reference bias and show that vg improves mapping of RNA-seq data compared to other mapping algorithms. We also demonstrate that our novel method, rpvg, can accurately estimate expression among millions of haplotype-specific transcripts derived from the GENCODE transcript annotation and the haplotypes from the 1000 Genomes Project.

## *Long Read Sequence-to-Graph Alignment using A* Seed Heuristic*

Pesho Ivanov (ETH Zurich) and Martin Vechev (ETH Zurich).

We present a practical algorithm for optimal alignment of long noisy sequences to reference genome graphs. The alignments are optimized according to edit distance metric with specified non-negative costs. We follow a principled approach of phrasing the alignment problem as a shortest path problem. To solve it, we developed a novel A* heuristic that extracts seeds from the query, approximately aligns them to the reference graph, and computes a sparse function specific to the read to be aligned. We developed and implemented an efficient algorithm that precomputes the heuristic and queries it in constant time. We evaluate the performance of the proposed algorithm on long noisy reads and compare it with the performance of approximate state-of-the-art long read aligners.

## *Using long and short reads to investigate interfering RNAs in coronavirus*

James Kelley (Rutgers University) and Andrey Grigoriev (Rutgers University).

see long abstract

## *Cell trajectory inference for revealing differentiation process of mouse stem cells using single cell RNA-seq data*

Takumi Adachi (Department of Bioinformatic Engineering, Osaka University), Junko Yoshida (Department of Physiology II, Nara Medical University), Shigeto Seno (Department of Bioinformatic Engineering, Osaka University), Kyoji Horie (Department of Physiology II, Nara Medical University) and Hideo Matsuda (Department of Bioinformatic Engineering, Osaka University).

Single-cell RNA-seq analysis is an effective tool for revealing differences in various cell types and states. Analyzing gene expression data is considered to be an effective approach for discovering biological mechanisms, especially cellular heterogeneity. In particular, understanding how cells change under different conditions is a very important issue for the further effective use of stem cells. However, a robust approach to express the process of state change based on gene expression data has not been established. In this presentation, we propose a method for identifying cell lineages for a data set containing cells in two different conditions: mutant and wild type. First, cluster cells based on the gene expression level, then infer a cell trajectory by applying partition-based graph abstraction (PAGA). We used Seurat version 3.1 and Scanpy version 1.4.4 for clustering and constructing PAGA graphs. We applied our method to the trajectory inference of the differentiation process of mouse embryonic stem cells with two time points, days 0 and 5 . We will present the results of comparing the trajectories between mutant and wild-type cells, and identify the effect of the mutation in the differentiation process.

## *Alignment of complex single-cell trajectories with CAPITAL*

Yuki Kato (Osaka University), Reiichi Sugihara (Osaka University), Tomoya Mori (Kyoto University) and Yukio Kawahara (Osaka University).

Recent techniques on single-cell RNA sequencing have boosted transcriptome-wide observation of gene expression dynamics of a heterogeneous cell population at a single-cell scale. Typical examples of such analysis include inference of a pseudotime cell trajectory, and comparison of pseudotime trajectories between different experimental conditions will tell us how feature genes regulate a dynamic cellular process. Existing methods for comparing pseudotime trajectories, however, force users to select trajectories to be compared because most of them can deal only with simple linear trajectories, leading to the possibility of making a biased interpretation. Here we present CAPITAL, a method for comparing complex pseudotime trajectories with tree alignment whereby trajectories including branching can be compared without any knowledge of paths in the trajectories to be compared. Computational tests on public data indicate that CAPITAL can align non-linear pseudotime trajectories at a sufficient level of known annotations, and reveal expression dynamics of several marker genes.

## *C4S DB: a database towards comprehensive collection and comparison for public ChIP-seq data*

Hayato Anzawa (Grad. Sch. Info. Sci., Tohoku Univ.) and Kengo Kinoshita (Grad. Sch. Info. Sci., Tohoku Univ.).

Although a massive ChIP-seq data is publicly available, current ChIP-seq databases have been established with a limited number of samples or insufficient quality assessment due to difficulty in large scale public ChIP-seq data analysis. To aim at providing ChIP-seq analysis results for comprehensive public data with quality assessable information, we launched C4S DB, a Comprehensive Collection and Comparison for ChIP-Seq database. We designed C4S DB to treat human ChIP-seq data released at the ENCODE portal and GEO. To cover as many as possible public datasets, we developed a semi-automated workflow to extract metadata for GEO ChIP-seq data. Since making researchers possible to choose data with suitable quality is crucial, our analysis pipeline includes steps to calculate several QC metrics which are comparable to the ENCODE project. Moreover, we also provide our novel QC metric, which does not require peak calling to evaluate ChIP noise level. C4S DB shares these statistics. It also provides the functions to visualise the orchestration of transcriptional regulatory elements around a gene and to map the similarity between experiments to infer regulatory elements' relationships. C4S DB will be a powerful platform to work with large scale public ChIP-seq data.

## *Dualrnaseq: a Nextflow-based workflow for host-pathogen Dual RNA-seq analysis.*

Bozena Mika-Gospodorz (Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI)) and Lars Barquist (Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI)).

RNA sequencing has become a standard technique for investigating the transcriptome of eukaryotes, bacteria, and archaea. The Dual RNA-seq protocol extends the RNA-seq technique to measure transcript expression in a host-pathogen system. However, simultaneous profiling of the gene expression of a pathogen and its host raises additional complications, particularly in read quantification.

Here, we present a Nextflow pipeline for quantifying Dual RNA-seq data. Host and pathogen RNA-seq reads are processed together, and, after quality control and trimming, are mapped onto two different references, the bacterial and eukaryotic genomes. We have implemented two read quantification strategies. The first is alignment-based mapping of reads onto the genomes with STAR followed by quantification with HTSeq, estimating gene expression using uniquely mapped reads. For the second we used Salmon with Selective Alignment, a fast transcriptome quantification method that handles multi-mapped reads. This allows us to investigate the importance of multi-mapped reads that may originate from different gene isoforms in the host and repetitive elements that are highly abundant in some bacterial genomes, as well as cross-mapping reads that may originate from either transcriptome. Using simulations of a variety of host-pathogen systems, we provide initial guidance towards optimal read quantification strategies for Dual RNA-seq experiments.

## *A novel probabilistic framework for exploring diversity in amplicon sequences with unique molecular identifiers*

Xiyu Peng (Iowa State University) and Karin Dorman (Iowa State University).

Amplicon sequencing has now been widely applied to explore genetic heterogeneity in populations and identify rare somatic mutations. The main challenge is how to confidently distinguish rare biological variants from noise sequences that can ultimately confound genetic analysis. One solution attaches Unique Molecular Identifiers (UMIs) to sample sequences before Polymerase Chain Reaction (PCR) amplification to increase the sensitivity for identifying ultra-low frequency mutations. Current algorithms that take into account UMI information either ignore errors within UMIs or cluster UMIs with an arbitrary threshold on the number of tolerated errors. Here we introduce a novel probabilistic framework for clustering amplicon sequences with unique molecular identifiers. A generative model is proposed not only for identifying biological variants but also for deduplicating PCR replicates. We test our model on both simulated data and a real Human Immunodeficiency Virus (HIV) dataset. The results show that our model achieves better accuracy than competing methods in identifying sequence variants and abundance estimation. Moreover, we show that clustering with UMIs can help correct possible biases in sequencing experiments.

## *Pannopi: a tool for pangenome based prokaryotic genome assembly and annotation*

Danil Zilov (Applied Genomics Laboratory, SCAMT Institute, ITMO University, Saint Petersburg) and Aleksey Komissarov (Applied Genomics Laboratory, SCAMT Institute, ITMO University, Saint Petersburg).

The advent of new generation sequencing for the first time made it possible to significantly speed up and reduce the cost of determining the full sequence of millions of genomes of organisms, from bacteria to humans. It is clear from the bioinformatics side that as the cost of sequencing decreases, the amount of data to be processed will increase. It is necessary to identify areas of analysis that are routine and to automate them. We created Pannopi - an open-source, scalable, easy-to-install, and use an assembly and annotation pipeline based on a hierarchical pan-genome graph. The program performs a large-scale analysis of the nucleic acid sequence of bacteria from preparation to functional annotation. The process runs from the preparation of sequence reads to genome assembly, through cleaning up the genome from external contamination to structural and functional annotation. Quality control is carried out throughout the process. Pannopi allows users to select the taxons for pan-genome graph construction and required modules; it can be used on a separate command-line program or through a web interface.